Int. J. on Recent Trends in Engineering and Technology, Vol. 8, No. 1, Jan 2013

A Simple Information Retrieval Technique

Sharifa Rania Mahmud
Military Institute of Science and Technology, Dhaka, Bangladesh
Email: sraniamahmud@gmail.com

***Abstract*—This research examines and analyzes the information retrieval techniques. The amount of information available over networks grows every day. This information worths being accessed and structured. Indexation and information retrieval are essential tasks to realize these objectives. This paper proposes an information retrieval technique which can retrieve appropriate document among a lot of documents. For doing this first, simplify all the documents. Then remove stop words and punctuations. It also calculates the term frequency, inverse term frequency, weight of each term etc. Here the proposed technique constructs the master document matrix. By this information retrieval technique anyone can easily search expected document from a collection of documents.**

***Index Terms*—Information Retrieval Techniques, Stop Words, Term Frequency, Master Document, Master Document Matrix**

I. Introduction

Information retrieval is the activity of obtaining information resources relevant to an information need from a collection of information resources. Searches can be based on metadata or on full-text indexing. The timely provision of relevant information with minimal *noise* is critical to modern society and this is what information retrieval (IR) is all about. It is a dynamic subject, with current changes driven by the expansion of the World Wide Web, the advent of modern and inexpensive graphical user interfaces and the development of reliable and low-cost mass storage devices. IR deals with the representation, storage, organization of, and access to information items [1].

Clearly, full description of the user information need cannot be used directly to request information using the current interfaces of Web search engines. Instead, the user must first translate this information need into a *query* which can be processed by the search engine. Given the user query, the key goal of an IR system is to retrieve information which might be useful or relevant to the user. IR technique consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need. There is overlap in the usage of the terms data retrieval, document retrieval, information retrieval, and text retrieval, but each also has its own body of literature, theory, praxis and technologies. IR is the study of systems for indexing, searching, and recalling data, particularly text or other unstructured forms [2, 8]. For an information retrieval system, the retrieved objects might be inaccurate and small errors are likely to go unnoticed. The main reason for this is that information retrieval usually deals with natural language text which is not always well structured and could be semantically ambiguous. To be effective in its attempt to satisfy the user information need, the IR system must somehow interpret the contents of the information items (documents) in a collection and rank them according to a degree of relevance to the user query. This interpretation of document content involves extracting syntactic and semantic information from the document text, and using this information to match the user information need. The difficulty is not only knowing how to extract this information but also knowing how to use it to decide relevance. In fact, the primary goal of an IR system is to retrieve all the documents which are relevant to a user query while retrieving as few non-relevant documents as possible [1].The purpose of this paper is to study the techniques of information retrieve and develop an easy and efficient information retrieval technique with which anyone can easily retrieve the expected document from a collection of documents.

A. Information Retrieval at the Center of the Stage

In the past 30 years, the area of information retrieval has grown well beyond its primary goals of indexing text and searching for useful documents in a collection. Nowadays, research in IR includes modeling, document classification and categorization, systems architecture, user interfaces, data visualization, filtering, languages, etc. Despite its maturity, until recently, IR was seen as a narrow area of interest. In the beginning of the 1990s, a single fact changed once and for all these perceptions — the introduction of the World Wide Web. The Web is becoming a universal repository of human knowledge and culture which has allowed unprecedented sharing of ideas and information in a scale never seen before. As a result, almost overnight, IR has gained a place with other technologies at the center of the stage [1].

More than 65 different technologies are recently available for web search. Among them the novel techniques are, Google, Yahoo!, Baidu, Bing, Yandex, Ask, AOL etc. Yahoo! and Google start their journey in 1995 and 1998 respectively. Google's worldwide market share peaked at 86.3% in April 2010. Yahoo!, Bing and other search engines are more popular in the US than in Europe. According to Hitwise, market share in the U.S. for October 2011 was Google 65.38%, Bing-powered (Bing and Yahoo!) 28.62%, and the remaining 66 search engines 6%. However, an Experian Hit wise report released in August 2011 gave the "success rate" of searches sampled in July. Over 80 percent of Yahoo! And Bing searches resulted in the users visiting a web site, while Google's rate was just under 68 percent.As research shows [13] people often come to IR systems with existing approaches to information seeking and processing and develop strategies for using specific systems. Many authors have pointed out

DOI: 01.IJRTET.8.1.53

that individual differences affect interaction with information and information systems [14, 15], that different stages of the search process require different kinds of assistance, and that differences in the search context affect the interactive support required-for example searching in secondary languages requires more support in the process of document assessment and querying [16].

II. Processing Of Documents

This section gives the details description of the proposed IR technique.

A. Stop Word

Some search engines don't record extremely common words in order to save space or to speed up searches. These are known as stop words. Stop words sometimes known as Noise Words (in the case of SQL Server) is the name given to words which are filtered out prior to, or after, processing of natural language data (text). Stop words ignored in a query because they are so commonly used that they can't contribute to relevancy. Includes conjunctions, prepositions, and articles such as and, to, a etc.

Hans Peter Luhn, one of the pioneers in information retrieval, is credited with coining the phrase and using the concept in his design. It is controlled by human input and not automated. This is sometimes seen as a negative approach to the natural articles of speech as mentioned above. There is no definite list of stop words which all natural language processing (NLP) tools incorporate. Not all NLP tools use a stop list. Some tools specifically avoid using them to support phrase searching. The use of a stemming algorithm may reduce part of the rationale or dependence on a stop list to filter out words. Google ignores common words and characters, such as where and how, as well as certain single digits and single letters [3, 4].

B. Punctuation

Punctuation is everything in written language other than the actual letters or numbers, including punctuation marks, inter-word spaces and indentation. Punctuation marks are symbols that correspond to neither phonemes (sounds) of a language nor lexemes (words and phrases), but which serve to indicate the structure and organization of writing, as well as intonation and pauses to be observed when reading it aloud. In English, punctuation is vital to disambiguate the meaning of sentences.

For example, "woman, without her man, is nothing", and "woman: without her, man is nothing", have greatly different meanings, as do "eats shoots and leaves" and "eats, shoots and leaves". The rules of punctuation vary with language, location, register and time, and are constantly evolving. Certain aspects of punctuation are stylistic and are thus the author's (or editor's) choice. Tachygraphic language forms, such as those used in online chat and text messages, may have wildly different rules [5].To simplify a document in information retrieval techniques stop words and punctuations have to delete from that document.

C. Term Frequency (TF)

Number of a word in a document is the frequency of that particular word in that document. If frequency of a word divided by the number of total words in a document, then the normalized frequency will be found. Sometimes it also called term frequency. The term frequency in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents (which may have a higher term frequency regardless of the actual importance of that term in the document) to give a measure of the importance of the term t_i within the particular document d_j.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

Where $n_{i,j}$ is the number of occurrences of the considered term in document d_j, and the denominator is the number of occurrences of all terms in document d_j [6].

D. Inverse Document Frequency (IDF)

The inverse document frequency is a measure of the general importance of the term (obtained by dividing the number of all documents by the number of documents containing the term, and then taking the logarithm of that quotient).

$$idf_i = log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \quad (2)$$

Where,

- $|D|$, total number of documents in the corpus
- $|\{d_j : t_i \in d_j\}|$, Number of documents where the term t_i appears, and $n_{i,j} \neq 0$ [7].

E. Weight of Each Word

Now weight of each word is:

$$tfidf_{i,j} = tf_{i,j} . idf_i \quad (3)$$

A high weight in tfidf is reached by a high term frequency (in the given document) and a low document frequency of the term in the whole collection of documents; the weights hence tend to filter out common terms. For example, consider a document containing 100 words wherein the word cow appears 3 times. By following the previously defined formulas, the TF for cow is, 0.03 (3/100). Now, assume 10 million documents are in database and cow appears in one thousand of these. Then, the inverse document frequency is calculated as *(10,000,000/1,000) = 9.21*. The *tfidf* or TF-IDF score is the product of these quantities: *(0.03*9.21) = 0.28* [6].

F. Master Document and Master Document Matrix

Master document is simply a list of words which has collection of all words from all documents. There are no repetitions of word in the master document. In this paper the proposed technique construct master document by taking union of terms or words from each document of the database and then sort the terms alphabetically.Master document matrix

DOI: 01.IJRTET.8.1.53

(MDM) is a matrix where weight of each term is placed. Terms are placed in rows and documents are placed in columns in the matrix. It is a large grid representing every documents and content words in a collection. It is also called term-document matrix (TDM). Generate MDM by arranging the list of all content words along the vertical axis, and a similar list of all documents along the horizontal axis [8].

For example, let there are three documents in the database, d1, d2 and d3. Documents are shown below:

- d1: Shipment of gold damaged in a fire.
- d2: Delivery of silver arrived in a silver truck.
- d3: Shipment of gold arrived in a truck.

Here in this example for simplicity stop words are not ignored and unique terms are sorted alphabetically [9]. Fig. 1 has shown the corresponding MDM and query matrix.

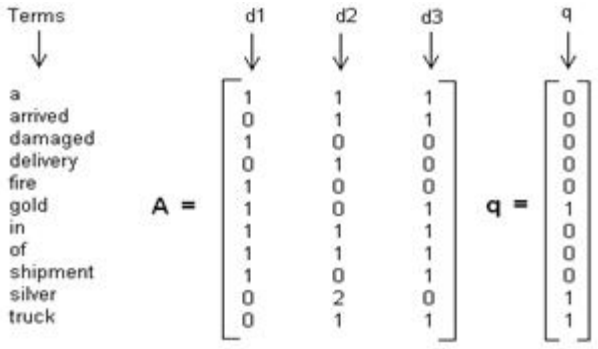

Fig. 1. Master Document Matrix and Query Matrix example.

G.. Ranking Documents

A ranking is a listing of items in a group, such as schools or sports teams, according to a system of rating or a record of performance [10]. Ranking document is giving a position of a document according to their similarity with a given query. Rank the document in first position which matches best with the query.

For example, suppose a query is q and the master document constructed is D.

q = [1 0 0 1 0 0 …..]

Here, 1 indicates that the term exists in the query, otherwise 0. Query-document similarity measures are then possible using:

sim(q, D) = q*D = [s1 s2 s3 s4 s5…….]

If, for example: s3>s1>s2>s5>s4, then the ranking of the documents will be:

d3>d1>d2>d5>d4

or, document-3> document-1> document-2> document-5> document-4

This paper proposed and develops an information retrieval technique. The proposed total algorithm for information retrieval technique works as in the following steps:

Algorithm:

Step 1: Stop words and punctuation remove from each document

Step 2: Calculate normalized frequency/term frequency of each term of each document

Step 3: Discard multiple occurrences of words and sort

Step 4: Calculate IDF

Step 5: Calculate weight of each word

Step 6: Construct master document by taking union of terms from each document

Step 7: Fill up the rows and columns of the master document matrix

Step 8: Similarity check and ranking the documents

III. Simulation Results

Suppose, for simulating the proposed algorithm take five documents for example. For simplicity very small documents are given as follows.

Document-1: Information retrieval is a wide, often loosely-defined term but in these pages I shall be concerned only with automatic information retrieval systems. Automatic as opposed to manual and information as opposed to data or fact.

Document-2: Information retrieval is the science of searching for documents, for information within documents and for metadata about documents, as well as that of searching relational databases and the World Wide Web.

Document-3: Stop words sometimes known as stopwords or Noise Words is the name given to words which are filtered out prior to, or after, processing of natural language data. Hans Peter Luhn, one of the pioneers in information retrieval, is credited with coining the phrase and using the concept in his design.

Document-4: The term frequency in the given document is simply the number of times a given term appears in that document. This count is usually normalized to prevent a bias towards longer documents to give a measure of the importance of the term within the particular document.

Document-5: Data retrieval, in the context of an IR system, consists mainly of determining which documents of a collection contain the keywords in the user query which, most frequently, is not enough to satisfy the user information need. In fact, the user of an IR system is concerned more with retrieving information about a subject than with retrieving data which satisfies a given query.

Using normal PHP program the simulation code has been done. In this paper, the simulation code has not shown. The final MDM is shown below in Fig. 2.Then we check the similarity and rank the documents.

For example, Let, the searching query is: “information retrieval system”

Then, *q=[0 0 0 0 0 1 .. 1 .. 1 ..]*

*sim(q,D) = q*D*

= [0.6319 0.3964 0.3052 0 0.8567]

So, the ranking is: *doc-5> doc-1> doc-2 > doc-3> doc-4*

There are various advantages of the proposed information retrieval technique. Save time to search information, helps in finding out the appropriate information, increases efficiency of work. There are many searching techniques available such as– genetic algorithm, hibernate, stemming algorithm, string searching algorithm etc. They are also very useful techniques but the problem is these algorithms are very hard to implement. On the other hand our searching algorithm is very easy to

DOI: 01.IJRTET.8.1.53

Docs / Words	doc-1	doc-2	doc-3	doc-4	doc-5
Appears	0	0	0	.02795	0
automatic	.07357	0	0	0	0
bias	0	0	0	.02795	0
coining	0	.02410	0	0	0
data	.011675	0	.00765	0	.013443
information	.01530	.01211	.00334	0	.005872
. .					
retrieval	.010120	.00605	.00334	0	.020165
. .					
system . .	0	0	0	0	.09794

Fig. 2. The Final Master Document Matrix.

understand and implement. So anyone can easily adopt and use this proposed technique.

Conclusion and Future Works

The proposed technique can be extended in future. We can use latent semantic indexing (LSI) and singular value decomposition (SVD). By using this technique execution time can be reduced. Latent Semantic Indexing tries to overcome the problems of lexical matching by using statistically derived conceptual indices instead of individual words for retrieval. A truncated singular value decomposition (SVD) is used to estimate the structure in word usage across documents. Retrieval is then performed using the database of singular values and vectors obtained from the truncated SVD. A number of software tools have been developed to perform operations such as parsing document texts, creating a term by document matrix, computing the truncated SVD of this matrix, creating the LSI database of singular values and vectors for retrieval, matching user queries to documents, and adding new terms or documents to an existing LSI databases. The bulk of LSI processing time is spent in computing the truncated SVD of the large sparse term by document matrices [11].

The paper develops an easy and efficient way to retrieve information from a collection of information. The paper introduces with the terms- stop word, punctuation, term frequency, IDF etc. Here the proposed technique simplify each document by removing stop words, punctuations and weighted each term of each document, by which appropriate documents are searched. This information retrieval technique can easily search the appropriate documents the user searching for. User gives a query for searching documents. That query can match with a lot of documents. Algorithm searches all the documents which match with the query. The result is a list of documents where documents are ranked. The document which matches best with the query is ranked as first, then the second and so on.

References

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto , "Modern Information Retrieval," ©Addison Wesley Longman Publishing Co. Inc., 1999.
[2] Information retrieval. Available from URL, http://en.wikipedia.org/wiki/Information_retrieval
[3] Stop word. Available from URL,http://en.wikipedia.org/wiki/Stop_words
[4] Danny Sullivan, "Search Engine Watch," Jan 1, 2003
[5] Punctuation. Available from URL,http://en.wikipedia.org/wiki/Punctuation
[6] Term frequency-Inverse document frequency. Available from URL, http://en.wikipedia.org/wiki/Tf-idf
[7] J. Kaur and V. Gupta, "Effective Approaches for Extraction of Keywords," International Journal of Computer Science Issues, vol. 7, Issue 6, November 2010.
[8] Master document matrix. Available from URL,http://www.seobook.com/lsi/tdm.htm
[9] SVD and LSI Tutorial 4: Latent Semantic Indexing (LSI) How-to Calculations.
[10] Dr. E. Garcia, Mi Islita.com. Available from URL:http://www.miislita.com/information-retrieval-tutorial/svd-lsi-tutorial-4-lsi-how-to-calculations.html#term-document
[11] Ranking document. Available from URL,http://www.answers.com/ranking&r=67
[12] M.W. Berry, S.T. Dumais and G.W. O'Brien, "Using Linear algebra for Intelligent Information Retrieval," SIAM Review, vol. 37, No. 4, pp. 573-595, December 1995.
[13] K. S. Kim, B. Allan, "Cognitive and task influences on Web searching behavior," Journal of the American Society for Information Science and Technology, 43, 2002, pp. 109–119.
[14] N. Ford, D. Miller and N. Moss, "Web search strategies and human individual differences: A combined analysis," Journal of the American Society for Information Science and Technology, 56, 2005, pp. 757–764.
[15] D. J. Slone, "The influence of mental models and goals on search patterns during Web interaction," Journal of the American Society for Information Science and Technology, 53, 2002, pp. 1152–1169.
[16] P. Hansen and J. Karlgren, "Effects of foreign language and task scenario on relevance assessment," Journal of Documentation, 61, 2005, pp. 623–639.

DOI: 01.IJRTET.8.1.53